FUZZY LOGISTIC REGRESSION: A NEW POSSIBILISTIC MODEL AND ITS APPLICATION IN CLINICAL VAGUE STATUS

S. POURAHMAD, S. M. T. AYATOLLAHI AND S. M. TAHERI

Abstract. Logistic regression models are frequently used in clinical research and particularly for modeling disease status and patient survival. In practice, clinical studies have several limitations For instance, in the study of rare diseases or due ethical considerations, we can only have small sample sizes. In addition, the lack of suitable and advanced measuring instruments lead to non-precise observations and disagreements among scientists in defining disease criteria have led to vague diagnosis. Also, specialists often report their opinion in linguistic terms rather than numerically. Usually, because of these limitations, the assumptions of the statistical model do not hold and hence their use is questionable. We therefore need to develop new methods for modeling and analyzing the problem.

In this study, a model called the " fuzzy logistic model " is proposed for the case when the explanatory variables are crisp and the value of the binary response variable is reported as a number between zero and one (indicating the possibility of having the property). In this regard, the concept of " possibilistic odds " is also introduced. Then, the methodology and formulation of this model is explained in detail and a linear programming approach is use to estimate the model parameters. Some goodness-of-fit criteria are proposed and a numerical example is given as an example.

1. Introduction

In classical set theory, an element either belongs to a set or it does not. In other words, the status of the element relative to the set is obvious. This property is related to the definition of that set. Some sets , such as the set of natural numbers equal or greater than 10 and the blood groups of the people, have such well-defined and precise criteria that people with varied levels of knowledge are in total agreement regarding their members.

Now, consider the set of *tall* men, the set of real numbers *much greater* than 10 or the set of patients with *high* blood pressure. There may be disagreement on the elements of these sets, because their definition contains imprecise or vague language; for example the word "tall" may mean different things to different people. These imprecisely defined sets, called "Fuzzy Sets", [25] play an important

Received: September 2009; Revised: January 2010 and February 2010; Accepted: March 2010
Key words and phrases: Logistic regression, Clinical research, Fuzzy logistic regression, Possibilistic odds.

role in human thought. In many scientific researches, linguistic, rather than numerical terms are frequently used. For instance, in clinical research, to measure the severity of disease or pain in patients, linguistic terms like *low, medium, high, very high,...* are used. These terms can be viewed as fuzzy sets. Moreover, the borderline between these fuzzy sets is not crisp even if they are measured in numerical scale. For example, to define *high* blood glucose in determining diabetic patients, cut-off point of 140 $[milligram(mg)]/[decilitre(dl)]$ [8] for two-hour plasma glucose during an oral glucose tolerance test is not the exact borderline. In other words, cases in the neighborhood of the borderline have a vague status with regard to the disease. A similar situation occurs in the definition of hypertension. To model the relationship between these observations, an ordinary statistical model which is based on certain assumptions and exact observations is not a good choice. For instance, in ordinary regression models, by the use of exact observations and based on some probability assumptions (such as normality and identically distribution of error terms), a function is built to predict the dependent variable from independent variables. Although statistical linear regression has many applications, problems can arise for small data sets, non-normality of the error terms, vagueness in the relationship between variables, ambiguity associated with the observed data, and inappropriateness of linearity assumption [19]. These are the situations in which fuzzy regression methods are applicable.

We note that fuzzy models as compared to the usual statistical models consider "possibilistic" rather than "probabilistic" errors. In other words, there are some aspects of uncertainty that measure the vagueness of the phenomena (due to their inconsistency to the existent criteria or the vagueness in their definition) and cannot be summarized in random terms. This kind of uncertainty is evaluated by a measure called possibility. Hence error terms are deleted in fuzzy regression models and are, in fact, hidden in the fuzzy coefficients.

In the last decades, two main approaches to fuzzy regression were introduced and investigated. The first one, which is called possibilistic regression, was proposed by Tanaka et al. [23]. Another approach is the fuzzy least squares approach (based on the distance between fuzzy observations and fuzzy estimations/predictions) which was proposed by Diamond [6] and Celmins [4], simultaneously. These approaches have been revisited and developed by many researchers (see e.g. [1, 5, 10, 11, 22]). Both approaches are used in different practical researches such as forecasting models in economics [16], modelling temporal systems [17], and finding pedomodels in soil science and environmental researches [15]. For more information about fuzzy regression methods, we refer the reader to the study by Taheri [21].

Almost all previous studies on fuzzy regression have focused on linear models and nonlinear models have been seldom considered. A nonlinear regression model which is widely used in research, especially in classical clinical studies, is the logistic regression model. This method is particularly appropriate for models involving disease state (disease/healthy), patient survival (alive/dead), and decision making (yes/no) [20].

In practice, there are many situations in which the ordinary logistic regression method cannot be used. For example, due to lack of suitable instruments or well-defined and wholly accepted criteria for some diseases in clinical research, scientists frequently encounter non-precise observations. In this situation, the variations of the model error terms cannot be attributed wholly to the randomness of the phenomenon. Furthermore, the probabilistic assumptions of the logistic model are not fulfilled. To deal with such situations, we combine the logistic regression model with fuzzy set theory to represent a new model which we call a " fuzzy logistic regression model ".

2. **Fuzzy Logistic Regression Model**

As it is known, a linear regression model cannot be used to regress a binary response variable (variable with two categories) on a set of explanatory variables $X = (x_1, x_2, ..., x_n)$, because the distribution of the response variable is Bernoulli and its mean is a probability (p) in [0,1] interval. Obviously, no one can guarantee that a linear combination of explanatory variables is also in this interval. So, for a sample of m observations, a function of mean response (named " *logit* " function) is modelled as a linear combination of explanatory variables [1]. Since $y_i = 0,\ 1$, with $E(Y_i) = P(Y_i = 1) = p_i, 0 < p_i < 1$, by considering the logit function as the linear combination of interest, we have $g(p_i) = \ln(p_i/(1-p_i)) = X_i'B$. Therefore, we can write

$$E(Y_i) = p_i = \exp(X_i'B)/(1 + \exp(X_i'B)), i = 1, 2, ..., m. \tag{1}$$

where, B is the vector of slope parameters. The expression $p_i/(1-p_i)$ is called the " *odds* " of characteristic 1.

In a logistic regression model, the parameters are usually estimated by the maximum likelihood method and the importance of each explanatory variable is assessed by carrying out statistical tests of significance of the coefficients. The usual test statistics for this purpose are the Wald statistic and likelihood ratio test statistic. Usually, the overall goodness of the fit of the model is also tested using χ^2 statistics for models with one explanatory variable and the Hosmer-Lemeshow test for more than one explanatory variable. The ability of the model to discriminate between the two groups defined by the response variable is proposed as the area under the receiver operating characteristic curve (ROC) [3].

Logistic regression, like other statistical models, depends heavily on the following assumptions : 1) distribution assumptions (Bernoulli probability distribution for the response variable, uncorrelated explanatory variables, independent and identically distributed error terms ...), 2) adequate sample size (low power of model validity methods, with small sample sizes to detect deviations from the logistic), and 3) exact observations. These assumptions impose some limitations in practice. For example, in clinical studies, based on a set of known risk factors (independent variables), one can predict the status of the disease and the likelihood of its occurrence in future by use of logistic regression analysis. However, due to lack of suitable instruments or well-defined criteria, physicians may suspect their diagnosis and, therefore, cannot categorize a person in one of two response categories. So, the

binary response observations are non-precise and the relationship between variables is not as precise as required in ordinary logistic regression [14]. In other words, due to the vagueness of the response variable, we cannot calculate the probability of belonging to category 1 ($p = P(Y = 1)$) and model the probability odds ($p/(1-p)$). It should be considered that real observations of a binary response variable are 0 or 1 (every one has the property or not), but our knowledge is not enough to distinguish them. Usually, these non-precise observations are eliminated from analysis in the ordinary logistic regression model. But imagine a situation in which all of our observations of the binary responses are vague and hence we cannot consider any probability distribution for the response variable or model the relationship between the success probability and explanatory variables. This type of uncertainty is not related to randomness and probability. So, instead of the probability, we consider the other aspect of uncertainty (i.e. possibility) to model such observations. By consulting with an expert, each case is compared to the previously accepted criteria of category 1 elements and the degree of consistency with that category (possibility of belonging to category 1) is noted. Then the new statistical term named " *possibilistic odds* " is defined and modeled.

If the consistency degree with the known characteristic for each non-precise (fuzzy) case is represented by μ_i $(0 \leq \mu_i \leq 1)$, its complement is $(1 - \mu_i)$ $(0 \leq 1 - \mu_i \leq 1)$. Then the " *possibilistic odds* " is defined as follows:

Definition 2.1. For each fuzzy case, the ratio $\mu_i/(1 - \mu_i)$, which shows the possibility of having the considered property for the i-th case to not, is called the possibilistic odds.

2.1. **Methodology and Formulation.** Consider the data set $X_i = (x_{i1}, x_{i2}, ..., x_{in})$ $i = 1, 2, \ldots n$, where X_i is the vector of crisp observations on the independent variables like sex, age, marital status, weight, blood pressure and blood cholesterol for the i-th case. μ_i, the corresponding response observation, is a number in [0,1] and indicates the possibility of i-th case having the relevant property i.e. μ_i=$Poss(Y_i = 1)$. Therefore, the fuzzy logistic regression model with fuzzy coefficients, is as follows:

$$\widetilde{W}_i = \widetilde{b}_0 + \widetilde{b}_1 x_{i1} + ... + \widetilde{b}_n x_{in},\ i = 1, ..., m. \tag{2}$$

$\widetilde{b}_0, \widetilde{b}_1, ..., \widetilde{b}_n$ are the model parameters which are treated as fuzzy numbers and $\widetilde{W}_i = \ln(\mu_i/(1 - \mu_i))$ is the estimator of the logarithmic transformation of possibilistic odds. To simplify the calculation, we assume that the fuzzy numbers $\widetilde{b}_j = (a_j^c, s_j^L, s_j^R)_T,\ j = 1, ..., n$ are triangular (Appendix, Definition 5.2). Then, (Appendix, Propositions 5.4 and 5.5), $\widetilde{W}_i,\ i = 1, ..., m$ (fuzzy output) is the triangular fuzzy number. We have $\widetilde{W}_i = (f_i^c(x), f_{is}^L(x), f_{is}^R(x))_T$, where:

$$\begin{aligned} f_i^c(x) &= a_0^c + a_1^c x_{i1} + ... + a_n^c x_{in}, \\ f_{is}^L(x) &= s_0^L + s_1^L x_{i1} + ... + s_n^L x_{in}, \\ f_{is}^R(x) &= s_0^R + s_1^L x_{i1} + ... + s_n^R x_{in}. \end{aligned} \tag{3}$$

So, the membership function of the fuzzy estimated output can be shown as follows:

$$\widetilde{W}_i(w_i) = \begin{cases} 1 - \frac{f_i^c(x) - w_i}{f_{is}^L(x)}, & f_i^c(x) - f_{is}^L(x) \leq w_i \leq f_i^c(x), \\ \\ 1 - \frac{w_i - f_i^c(x)}{f_{is}^R(x)}, & f_i^c(x) < w_i \leq f_i^c(x) + f_{is}^R(x). \end{cases} \quad (4)$$

If $s_i^L = s_i^R = s_i,\ i = 1, ..., m$, the triangular fuzzy number is called symmetric. In this case, for $\widetilde{W}_i$ we have: $f_{is}^L(x) = f_{is}^R(x) = f_i^s(x)$.

As it is known, $\widetilde{W}_i$ is the natural logarithm of possibilistic odds of getting or having the known property for the i-th case. According to the extension principle (Appendix, Definition 5.3), if $\widetilde{M}$ is a fuzzy number with membership function $\widetilde{W}_i$, and $f(x) = \exp(x)$, then $f(\widetilde{M}) = \exp(\widetilde{M})$ is a fuzzy number with the following membership function:

$$\exp(\widetilde{M}(x)) = \begin{cases} \widetilde{M}(lnx), & x > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

So, after estimating the model coefficients, we can determine the membership function of the possibilistic odds $\exp(\widetilde{W}_i(x)),\ x > 0$ as follows:

$$\exp(\widetilde{W}_i(x)) = \widetilde{W}_i(\ln(x)) = \begin{cases} 1 - \frac{f_i^c(x) - \ln(x)}{f_{is}^L(x)}, & f_i^c(x) - f_{is}^L(x) \leq \ln(x) \leq f_i^c(x), \\ \\ 1 - \frac{\ln(x) - f_i^c(x)}{f_{is}^R(x)}, & f_i^c(x) < \ln(x) \leq f_i^c(x) + f_{is}^R(x). \end{cases} \quad (6)$$

Thus, for a new fuzzy case, our model can predict its possibilistic odds as a fuzzy number by the use of the crisp vector of input observations (explanatory observations),.

2.2. **Estimation of Fuzzy Coefficients.** In order to estimate the coefficients $\widetilde{b}_j,\ j = 0, 1, ..., m$, we follow the possibilistic approach [23] in fuzzy linear regression models with fuzzy output, fuzzy coefficients and non-fuzzy input vector. The basic idea is to minimize the fuzziness of the obtained model by minimizing the total support of the fuzzy coefficients. So, it is assumed that:

1) Each observation, w_i has a membership degree as big as h in the function of the fuzzy estimated output, $\widetilde{W}_i$, i.e.

$$\widetilde{W}_i(w_i) \geq h \ and \ w_i = \ln(\frac{\mu_i}{1 - \mu_i}),\ h \in (0, 1) \quad (7)$$

2) The fuzzy coefficients $(\widetilde{b}_j,\ j = 0, 1, ..., m)$ are such that the fuzziness of the model is minimized. Since the fuzziness of a fuzzy number increases with its spreads, minimizing the sum of the spreads of fuzzy outputs leads to a minimum value of the fuzziness of the model.

The determination of fuzzy coefficients leads to a linear programming problem, in which the objective function is the sum of the spreads of the fuzzy outputs,

$$Z = m(s_0^L + s_0^R) + \sum_{j=1}^{n}[(s_j^L + s_j^R)\sum_{i=1}^{m} x_{ij}] \quad (8)$$

where, x_{ij} is the value of the i-th observation for the j-th variable.

On the other hand, by equation 6 each constraint of the problem $(\widetilde{W}_i(w_i) \geq h,\ i = 1, ..., n)$ can be written as follows:

$$1 - \frac{f_i^c(x) - w_i}{f_{is}^L(x)} \geq h \Longrightarrow (1-h)s_0^L + (1-h)\sum_{j=1}^{n} s_j^L x_{ij} - a_0^c - \sum_{j=1}^{n} a_j^c x_{ij} \geq -w_i$$

$$1 - \frac{w_i - f_i^c(x)}{f_{is}^R(x)} \geq h \Longrightarrow (1-h)s_0^R + (1-h)\sum_{j=1}^{n} s_j^R x_{ij} + a_0^c + \sum_{j=1}^{n} a_j^c x_{ij} \geq w_i. \quad (9)$$

Hence we have a total of $2m$ constraints. One can minimize the objective function (equation 8) using linear programming algorithms, such as the Simplex method, to estimate the mode value, and the left and the right spreads of each coefficient. The GAMS [9], Lingo 8.0 [12], and Matlab [13] may be used.

2.3. **Goodness-of-fit Criteria.** Like other statistical modelling, the models based on fuzzy rules, need to be evaluated by some methods (how well the model fits the data). Several goodness-of-fit methods for fuzzy models have been proposed [18]. Here we propose two methods for the evaluation of fuzzy logistic regression models.

2.3.1. **Mean Degree of Memberships (MDM).** As mentioned earlier, for i-th case $i = 1, ..., m$, there are two values: the observed value of the response variable, w_i, which is a crisp number and the estimated value, $\widetilde{W}_i$, which is a fuzzy number (after modelling).

Definition 2.2. Consider the fuzzy logistic regression model which is derived based on m crisp observations. Then, the Mean Degree of Memberships (MDM) for observed values in the membership function of the estimated ones, i.e.

$$MDM = \frac{1}{m}\sum_{i=1}^{m} \widetilde{W}_i(w_i) = \frac{1}{m}\sum_{i=1}^{m} \exp(\widetilde{W}_i(\frac{\mu_i}{1-\mu_i})) \quad (10)$$

is used as an index for evaluating the model.

Large membership degrees of the observed values confirm that the model constructed from these data supports the data well. The maximum value of MDM index is 1 ($\frac{1}{m}\sum_{i=1}^{m} 1$) and the minimum value is 0 ($\frac{1}{m}\sum_{i=1}^{m} 0$). So, a value near 1 indicates good model fitting.

2.3.2. **Mean of Squares Errors (MSE).** Another method for evaluating the goodness of fit is to measure distance of two outputs from each other. The nearer the estimated output to the observed one the higher the power of the model to predict the real situation of the cases. However, since the observed response value is a crisp number while the estimated value is a fuzzy number, it is necessary that the fuzzy number should first be changed to a crisp number using defuzzification methods. [7].

Definition 2.3. Consider the fuzzy logistic regression model with crisp input observations, $\mu_i/(1-\mu_i)$, and fuzzy estimated outputs $\exp(\widetilde{W}_i)$. The MSE index of the model is defined as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}[def(\exp(\widetilde{W}_i)) - (\frac{\mu_i}{1-\mu_i})] \tag{11}$$

in which, $def(\exp(\widetilde{W}_i))$ is the defuzzification of $\exp(\widetilde{W}_i)$ (Appendix, Definition 5.6)

3. Application in Clinical Studies

In clinical diagnosis, the performance of all screening tests depends on the cut-off point used to separate normal and abnormal individuals. A too high cut-off point may cause abnormal individuals to be classified as normal and a too low cut-off point may classify healthy individuals as abnormal. As an example, consider diabetes mellitus (DM) which is a common metabolic disorder that shares the phenotype of hyperglycemia. Several distinct types of DM exist and are caused by a complex interaction of genetics and environmental factors [8]. There are no widely accepted or rigorously validated cut-off points for defining positive screening tests for diabetes in non-pregnant adults [20]. The latest suggested cut-off point for fasting plasma glucose is "less than 100 (mg/dl) for normal and higher than 126 (mg/dl) for abnormal cases" and for two-hour postprandial plasma glucose is "less than 140 (mg/dl) for normal and 200 (mg/dl) for abnormal cases" [20]. However, identical cut-off point in all texts does not guarantee the crispness. In other words, in the neighborhood of the cut-off point, a little increase or decrease in blood plasma glucose cannot change the individual's status from normal to abnormal. In this case, physicians do not rely on one result and repeat the examination to reach more reliability. Also, they consider any other clinical symptoms of diabetes (polyuria, polydipsia, weight loss, etc.) to decide the classification. However, they admit that the status of patients whose two-hour postprandial plasma glucose is in the 140-200 (mg/dl) interval or whosw fasting plasma glucose is in the interval 100-126 (mg/dl), is unknown (impaired glucose tolerance test). Fuzzicians believe that degree of vagueness in this interval is not the same. In the following example, we use the diabetes data to show the application of our proposed model. Obviously, this model can also be used in similar situations such as Behcet, systematic lupus erthematosus and hypertension .

3.1. A Numerical Example.

In order to determine the diabetic status of a community, in a clinical survey, a sample of two-hour postprandial plasma glucose of each person is available. Considering 200 (mg/dl) as cut-off point, it was found that 15 cases fell in the interval 140-200 (mg/dl). In order to predict the possibilistic odds of diabetes for these "vague status" cases, used additional information such as sex (female), age (year), BMI ($[weight(kg)]/\ [(height(m))^2]$), family history (close relatives such as father, mother, sister and brother), and two-hour plasma glucose (mg/dl) which are shown to be significant risk factors in diabetes [20] (Table 1). We consulted with an expert to assign a possibility of the disease to each case.

Id	Sex	Two-hour postprandial plasma glucose (mg/dl)	Age (year)	Family history	BMI (kg/m^2)	μ_i	$\frac{\mu_i}{1-\mu_i}$	$w_i = \ln(\frac{\mu_i}{1-\mu_i})$
1	1	145	40	0	24	0.10	0.11	-2.20
2	1	147	42	0	25	0.15	0.18	-1.73
3	0	150	45	1	21	0.35	0.54	-0.62
4	0	155	37	1	23	0.42	0.72	-0.32
5	0	157	59	1	25	0.49	0.96	-0.04
6	1	160	44	0	20	0.50	1.00	0.00
7	1	160	38	1	26	0.60	1.50	0.41
8	1	165	52	0	33	0.60	1.50	0.41
9	0	182	50	0	31	0.70	2.33	0.85
10	1	187	55	1	33	0.85	5.67	1.73
11	0	190	53	1	35	0.90	9.00	2.20
12	0	192	62	1	30	0.97	32.33	3.48
13	0	195	57	0	32	0.95	19.00	2.94
14	1	195	50	0	34	0.95	19.00	2.94
15	1	196	60	1	35	0.99	99.00	4.60

TABLE 1. Risk Factors and Membership Degrees of 15 Diabetic Fuzzy Cases

Then, the following possibilistic model was fitted:

$$\widetilde{W_i} = \widetilde{b}_0 + \widetilde{b}_1\ Sex_i + \widetilde{b}_2\ Blood\ glucose_i + \widetilde{b}_3\ Age_i + \widetilde{b}_4\ BMI_i + \widetilde{b}_5\ Family\ history_i \quad i = 1, ..., 15 \tag{12}$$

where, for simplicity in computations, the regression coefficients $\widetilde{b}_j = (a_j^c, s_j^L, s_j^R)_T$, $j = 0, 1, ..., 5$ are assuned to be triangular fuzzy numbers and the fuzziness of the variables' relationships is hidden in these coefficients. Depending on the definition of the coefficients in fuzzy models, there are two types of models:

a) The model with symmetric coefficients: $\widetilde{b}_j = (a_j^c, s_j^L, s_j^R)_T$ in which $s_j^L = s_j^R = s_j,\ j = 0, 1, ..., 5$.

b) The model with non-symmetric coefficients: $\widetilde{b}_j = (a_j^c, s_j^L, s_j^R)_T$ in which $s_j^L \neq s_j^R$, for some $j = 0, 1, ..., 5$.

Both models for different h values were fitted to our data and the results obtained were similar (calculation is not shown). So, for simplicity in computation and interpretation, we choose the symmetrical model to fit our data. Now, to decide about the h value (see equation 7), we fit the symmetrical model for several values of h and observe its effect on the model coefficients (Table 2.).

As shown, changing h values do not change the coefficient centers (a_j^c) but affects the spreads (s_j) and objective function (Z) values such that the vagueness of the fuzzy

h	s_0	s_1	s_2	s_3	s_4	s_5	a_0^c	a_1^c	a_2^c	a_3^c	a_4^c	a_5^c	$Z_{(a)}$
0.1	0.0	0.24	0.0	0.0	0.01	0.50	-15.88	0.49	0.09	0.07	-0.11	0.49	19.57
0.2	0.0	0.29	0.0	0.0	0.01	0.56	-15.88	0.49	0.09	0.07	-0.11	0.49	22.01
0.3	0.0	0.33	0.0	0.0	0.01	0.65	-15.88	0.49	0.09	0.07	-0.11	0.49	25.16
0.4	0.0	0.38	0.0	0.0	0.01	0.75	-15.88	0.49	0.09	0.07	-0.11	0.49	29.35
0.5	0.0	0.46	0.0	0.0	0.01	0.90	-15.88	0.49	0.09	0.07	-0.11	0.49	35.22
0.6	0.0	0.58	0.0	0.0	0.02	1.13	-15.88	0.49	0.09	0.07	-0.11	0.49	44.02
0.7	0.0	0.78	0.0	0.0	0.02	1.50	-15.88	0.49	0.09	0.07	-0.11	0.49	58.70
0.8	0.0	1.16	0.0	0.0	0.04	2.26	-15.88	0.49	0.09	0.07	-0.11	0.49	88.04
0.9	0.0	2.34	0.0	0.0	0.08	4.52	-15.88	0.49	0.09	0.07	-0.11	0.49	176.09

TABLE 2. The Objective Function and Coefficients of Symmetrical Model for Different Values of h

outputs increased with the h values. So, based on the Z values and the vagueness of the outputs, it seems that the value 0.6 is the rational choice for h. Now the symmetrical possibilistic logistic regression model with $h = 0.6$ is fitted to our data. As mentioned in Section 2.2, to fit the model by linear programming methods [23] and in order to determine the coefficients $\widetilde{b}_j,\ j = 0, 1, ..., n$, the objective function should be minimized in such a way that two constraints for each observation are satisfied. The objective function in our example is:

$$Z = 2(15\ s_0 + s_1 \sum_{i=1}^{15} Sex_i + s_2 \sum_{i=1}^{15} Blood\ glucose_i + s_3 \sum_{i=1}^{15} Age_i + s_4 \sum_{i=1}^{15} BMI_i + s_5 \sum_{i=1}^{15} Family\ history_i) = 2(15\ s_0 + 8\ s_1 + 2576\ s_2 + 744\ s_3 + 427\ s_4 + 8\ s_5) \tag{13}$$

This function should be minimized under 30 constraints (15 observations $\times$ 2). For instance, the two constraints of the first observation are as follows:

$$\begin{aligned} 0.1s_0 + 0.1s_1 + 14.5s_2 + 4s_3 + 2.4s_4 - a_0^c - a_1^c - 145a_2^c - 40a_3^c - 24a_4^c \geq -2.2, \\ 0.1s_0 + 0.1s_1 + 14.5s_2 + 4s_3 + 2.4s_4 + a_0^c + a_1^c + 145a_2^c + 40a_3^c + 24a_4^c \geq 2.2. \end{aligned} \tag{14}$$

Using the Lingo software [12], the above linear programming problem was solved and the coefficients estimated were as follows: $a_0^c = -15.8833,\ a_1^c = 0.4884,\ a_2^c = 0.0927,\ a_3^c = 0.0727,\ a_4^c = -0.1142,\ a_5^c = 0.4940,\ s_0 = 0.0,\ s_1 = 0.5841,\ s_2 = 0.0,\ s_3 = 0.0,\ s_4 = 0.0194,\ s_5 = 1.1305.$

The minimized value of the objective function was: $Z = 44.02$, and the optimal model was obtained as:

$$\begin{aligned} \widetilde{W} = -15.8833 + (0.4884,\ 0.5841)_T\ Sex + 0.0927\ Blood\ glucose + 0.0727\ Age \\ +(-0.1142,\ 0.0194)_T\ BMI + (0.4940,\ 1.1305)_T\ Family\ history, \end{aligned} \tag{15}$$

This formula can estimate the possibility odds of diabetes for a case that is suspected in a diabetic status. Note that, the estimated possibility odds for each case

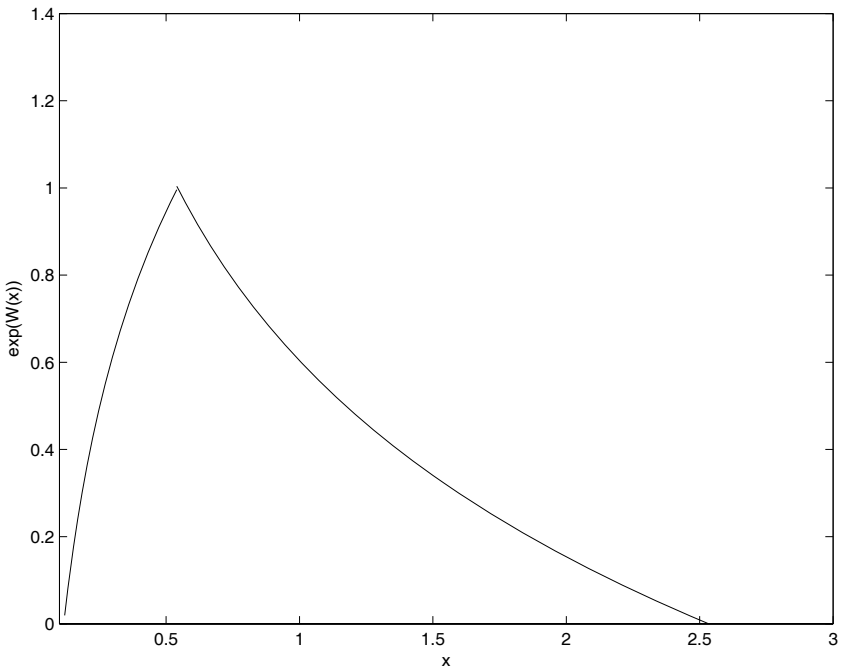

FIGURE 1. The Membership Function of *about* 0.54 (in equation 17)

is a fuzzy output. For instance, suppose we want to predict the possibility odds of disease for the case numbered 3 in Table 1. We have:

$$\begin{aligned}\widetilde{W}_3 = -15.8833 + (0.4884,\ 0.5841)_T \times 0 + 0.0927 \times 150 + 0.0727 \times 45 + (-0.1142,\ 1.1305)_T \\ \times 21 + (0.4940,\ 1.1305)_T \times 1 = (-15.8833,\ 0)_T + (13.9050,\ 0)_T + (3.2715,\ 0)_T + \\ (-2.3982,\ 0.4074)_T + (0.4940,\ 1.1305)_T = (-0.6110,\ 1.5379)_T = (-0.61,\ 1.54)_T\end{aligned} \tag{16}$$

This means that $\widetilde{W}_3 = (-0.61,\ 1.54)_T$, i.e., the logarithm of possibilistic odds for case 3 which is rounded to 2 decimals, is about -0.61 (a triangular fuzzy number). And now, to calculate the estimated possibilistic odds, as mentioned in Section 2.1, using the extension principle we have (Appendix, Definition 5.3):

$$\exp(\widetilde{W}_3(x)) = \begin{cases} \widetilde{W}_3(\ln x), & x > 0; \\ 0, & otherwise. \end{cases}$$

$$= \begin{cases} 1 - \frac{-0.61 - \ln x}{1.54}, & -2.15 \le \ln x \le -0.61 \ (0.12 \le x \le 0.54), \\ 1 - \frac{\ln x + 0.61}{1.54}, & -0.61 < \ln x \le 0.93 \ (0.54 < x \le 2.53). \end{cases} \tag{17}$$

This means that the possibilistic odds of diabetes of case 3 is "*about* 0.54". The membership function of this linguistic term is shown in Figure 1.

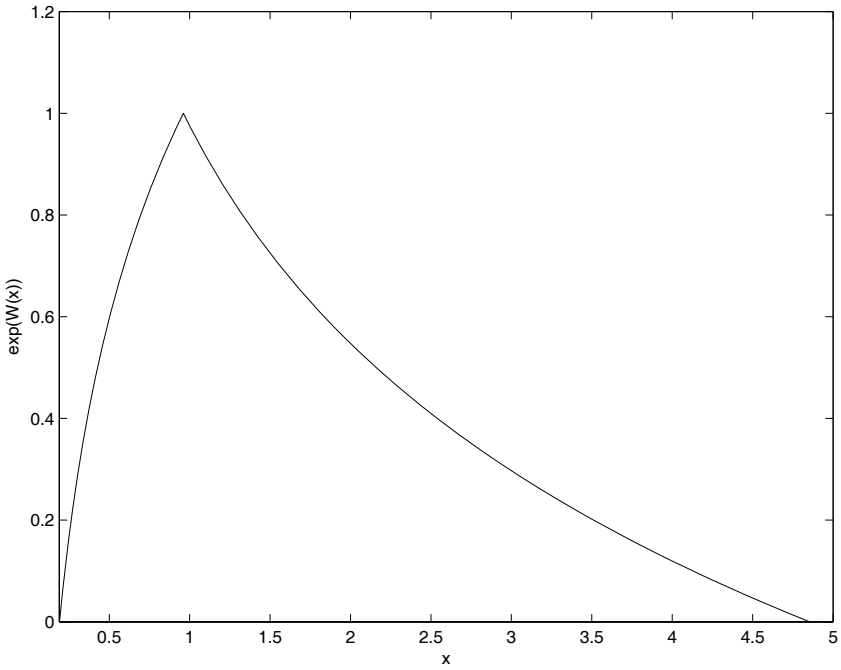

FIGURE 2. The Membership Function of *about* 0.96 (in equation 18)

Now, suppose there is a new person with ($x_1 = 0,\ x_2 = 165,\ x_3 = 40,\ x_4 = 25,\ x_5 = 0$). Then, according to model 15, the possibilistic odds of diabetes for this person is obtained as:

$$\widetilde{W}_{new} = -15.8833 + (0.4884,\ 0.5841)_T \times 0 + 0.0927 \times 165 + 0.0727 \times 40 + (-0.1142,\ 1.1305)_T$$
$$\times 25 + (0.4940,\ 1.1305)_T \times 1 = (-15.8833,\ 0)_T + (15.2955,\ 0)_T + (2.9080,\ 0)_T +$$
$$(-2.8550,\ 0.4850)_T + (0.4940,\ 1.1305)_T = (-0.0408,\ 1.6155)_T = (-0.0408,\ 1.6155)_T,$$

$$\exp(\widetilde{W}_{new}(x)) = \begin{cases} 1 - \frac{-0.04 - \ln x}{1.62}, & -1.66 \leq \ln x \leq -0.04 \ (0.19 \leq x \leq 0.96); \\ 1 - \frac{\ln x + 0.04}{1.62}, & -0.04 < \ln x \leq 1.58 \ (0.96 < x \leq 4.85). \end{cases} \tag{18}$$

So, we can say that, the possibilistic odds of diabetes for this new case is about 0.96 (Figure 2).

Finally, to evaluate the model, we use the two criteria proposed in Section 3, i.e., MDM and MSE. To find the MDM index, we calculate the membership degree of each observed odds in the membership function of its related fuzzy output (equation 10). The results are shown in Table 3.

Id	$\frac{\mu_i}{1-\mu_i}$	$w_i = \ln(\frac{\mu_i}{1-\mu_i})$	$\widetilde{W}_i$	$\exp(\widetilde{W}_i(\frac{\mu_i}{1-\mu_i})) = \widetilde{W}_i(w_i)$
1	0.11	-2.20	$(-1.79,\ 1.05)_T$	0.61
2	0.18	-1.73	$(-1.57,\ 1.07)_T$	0.85
3	0.54	-0.62	$(-0.61,\ 1.54)_T$	0.99
4	0.72	-0.32	$(-0.96,\ 1.58)_T$	0.60
5	0.96	-0.04	$(0.60,\ 1.62)_T$	0.60
6	1.00	0.00	$(0.35,\ 0.97)_T$	0.64
7	1.50	0.41	$(-0.28,\ 2.22)_T$	0.69
8	1.50	0.41	$(-0.09,\ 1.22)_T$	0.59
9	2.33	0.85	$(1.08,\ 0.60)_T$	0.61
10	5.67	1.73	$(2.66,\ 2.35)_T$	0.60
11	9.00	2.20	$(2.08,\ 1.81)_T$	0.93
12	32.33	3.48	$(3.49,\ 1.71)_T$	0.99
13	19.00	2.94	$(2.68,\ 0.62)_T$	0.59
14	19.00	2.94	$(2.43,\ 1.24)_T$	0.59
15	99.00	4.60	$(3.63,\ 2.39)_T$	0.60

TABLE 3. The Observed and the Estimated fuzzy Outputs of Diabetes Data

The last column of Table 3 is calculated assuming symmetrical triangular fuzzy numbers (Appendix, Definition 5.2) and the membership degree of $\mu_i/(1-\mu_i)$ is a crisp number in this function. Finally we obtain

$$MDM = \frac{1}{m}\sum_{i=1}^{m}\exp(\widetilde{W}_i(\frac{\mu_i}{1-\mu_i})) = \frac{10.48}{15} = 0.70 \tag{19}$$

The MDM value is much greater than 0.5 indicating a good fit.

To obtain the index MSE, we defuzzify each output and then calculate its distance to the corresponding observation (Table 4). For example, in case 6, by the Center of Gravity defuzzification method (Appendix, Definition 5.6), we have:

$$def_{CoG}(\exp(\widetilde{W}_6)) = \frac{\int_{\exp(0.35-0.97)}^{\exp(0.35)} x(1-\frac{0.35-\ln x}{0.97})dx + \int_{\exp(0.35)}^{\exp(0.35+0.97)} x(1-\frac{\ln x-0.35}{0.97})dx}{\int_{\exp(0.35-0.97)}^{\exp(0.35)}(1-\frac{0.35-\ln x}{0.97})dx + \int_{\exp(0.35)}^{\exp(0.35+0.97)}(1-\frac{\ln x-0.35}{0.97})dx} = 1.78 \tag{20}$$

Table 4 shows the results of calculations. The MSE value of the model is obtained as:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}[def_{CoG}(\exp(\widetilde{W}_i) - \frac{\mu_i}{1-\mu_i})]^2 = 12.91 \tag{21}$$

As mentioned before, this index shows the mean distance between the observed and the estimated response values. Its small value confirms a good fit. Unfortunately, there are not any critical values with which to compare our indices. It seems that the indices like MSE are useful when we are interested in comparing several

Id	$\frac{\mu_i}{1-\mu_i}$	$def_{CoG}(\exp(\widetilde{W}_i))$	$\lvert def_{CoG}[\exp(\widetilde{W}_i)] - \frac{\mu_i}{1-\mu_i}\rvert$
1	0.11	0.22	0.11
2	0.18	0.27	0.09
3	0.54	0.93	0.39
4	0.72	0.68	0.04
5	0.96	3.30	2.34
6	1.00	1.78	0.78
7	1.50	2.14	0.64
8	1.50	1.30	0.20
9	2.33	3.22	0.89
10	5.67	4.96	0.71
11	9.00	16.56	7.56
12	32.33	36.19	3.86
13	19.00	16.03	2.97
14	19.00	16.31	2.69
15	99.00	108.88	9.88

TABLE 4. Distance Between the Defuzzified Estimated Possibilistic Odds and Its Related Observed One

fuzzy models for the same data set and choosing the best one. The model with the smallest MSE value is chosen.

4. **Discussion**

Usually, the real circumstances of the data do not completely match the underlying distributional assumptions of theoretical statistical models. This encourages researchers to model the data in a more flexible environment which is closer to the real situation of the observations and one choice is using fuzzy models. These models have been widely studied and applied in diverse areas. However, there is no doubt that fuzzy models are more complicated than ordinary ones in calculation and interpretation, but ordinary statistical models' use is limited to their assumptions. If the data do not satisfy the model assumptions, the use of ordinary methods is not rational and causes bias in the results. Note that, the ordinary models cannot be replaced by fuzzy models or vice versa because their applications are different from each other. Usually, these two models cannot be used for the same data set simultaneously and, consequently, their results cannot be compared with each other.

Non-precise observation is one of the situations in which fuzzy modelling methods is recommended. These observations are frequently seen in clinical studies. Sometimes, there are errors in clinical measuring instruments. Also, there are some ethical considerations in these studies. In such cases, the exact value of variables often cannot be measured and the observations are reported approximately. Another uncertain situation in clinical studies is in disease diagnosis, where a disease

is diagnosed according to the known criteria. If a person has all the disease symptoms, he/she is considered as a patient and whenever he/she does not have any symptoms, he/she is considered a healthy person. What happens when someone has only some of these symptoms? In this case the physician does not really know whether to start the therapy or not. Also, to determine a borderline (cut-off point) between patients and healthy people, there is no crisp cut-off point in clinical laboratory tests. It means that all the individuals near the cut-off point have fuzzy status. In order to determine the more important risk factors (the factors which advance the disease), these vague observations are usually not used in the ordinary modeling analysis, and discarding or ignoring them from analysis is not rational. It appears that fuzzy models are appropriate methods in this situation.

In the current paper, the details of a fuzzy logistic regression are discussed and a numerical example of its application to clinical studies is given. The proposed model is applied when the observations of the binary response variable are vague (i.e. instead of 0 or 1, they are reported as a value in [0,1] representing the possibility of having the disease) but the observations of the explanatory variables are precise.

Since the binary response variable has a vague status, the $P(Y = 1)$ is not definable and the probability odds cannot be calculated. The value of μ_i detects the degree of adjustment to the category 1 criteria of the response variable for i-th case (its possibility) and is determined by a clinical expert. In our model it is a number between 0 and 1. However, it can be a linguistic qualitative term such as *low*, *medium*, *high*, *very high*, (which is recommended for use in a new model). To estimate fuzzy coefficients in our proposed model, we used the possibilistic approach. In addition, to determine the aptness of the model, two indices were introduced, namely MDM (Mean Degree of Memberships) and MSE (Mean Square Errors).

The proposed model can also be used in other research areas with similar situations.

5. **Appendix**

We start this section with a simple definition of a *fuzzy set* and then point out some definitions and algebraic operations which are used in this paper. A *classical (crisp) set* is defined as a collection of elements or objects of a universal set. In such a classical set, we can relate a membership value to each element by using the characteristic function, in which 1 indicates membership and 0 non-membership. For a fuzzy set, the characteristic function allows various degrees of membership (partial membership) for the elements of the universal set. The definition of the membership function for a fuzzy set is more or less subjective but its values are always in the interval $[0, 1]$, indicating non-membership for 0, full membership for 1, and partial membership between 0 and 1 for the ambiguous elements. Typically, we have the following elementary definitions.

If X is a collection of objects denoted by x, then a fuzzy set $\widetilde{A}$ in X is defined as a set of ordered pairs: $\widetilde{A} = \{(x,\ \widetilde{A}(x))|x \in X\}$, where $0 \leq \widetilde{A} \leq 1$ is called the membership function representing the grade of membership of x in fuzzy set $\widetilde{A}$. The value $M = \sup_x \widetilde{A}(x)$ is called the height of $\widetilde{A}$. If $M = 1$, then $\widetilde{A}$ is

called normal. The complement set of fuzzy set $\widetilde{A}$, is a fuzzy set $\widetilde{B}$ with the membership function $\widetilde{B}(x) = 1 - \widetilde{A}(x)$. A fuzzy set $\widetilde{A}$ of $\Re$ is said to be convex if $\widetilde{A}(\lambda x_1 + (1-\lambda)x_2) \geq \min(\widetilde{A}(x_1), \widetilde{A}(x_2)),\ x_1, x_2 \in X,\ \lambda \in [0,1]$.

Definition 5.1. The fuzzy set $\widetilde{M}$ of the real line $\Re$ is called a fuzzy number if it is a normal convex fuzzy set of $\Re$.

Definition 5.2. A fuzzy number $\widetilde{N}$ is called of LR-type if it has the membership function as follows:

$$\widetilde{N}(x) = \begin{cases} L(\frac{m-x}{\alpha}), & x \leq m, \\ \\ R(\frac{x-m}{\beta}), & x > m. \end{cases} \quad ,\forall x \in \Re$$

where, L and R are decreasing shape functions from $\Re^+$ to $[0,1]$ with $L(0) = 1$, $L(x) < 1$ for all $x > 0$, $L(x) > 0$ for all $x < 1$ and $L(1) = 0$; (or $L(x) > 0$ for all x and $L(+\infty) = 0$). Similar conditions hold for R. The real number m is called the mean value of $\widetilde{N}$, and α and β (positive numbers) are called the left and the right spreads, respectively. Symbolically $\widetilde{N}$ is denoted by $(m, \alpha, \beta)_{LR}$. In the special case, where $L(x) = R(x)$, $\widetilde{N}$ is called triangular fuzzy number and is denoted by $(m, \alpha, \beta)_T$. Its membership function is:

$$\widetilde{N}(x) = \begin{cases} 1 - \frac{m-x}{\alpha}, & m - \alpha \leq x \leq m, \\ \\ 1 - \frac{x-m}{\beta}, & m < x \leq m + \beta. \end{cases} \quad ,\forall x \in \Re$$

If, in addition, $\alpha = \beta$, then $\widetilde{N}$ is denoted by $(m, \alpha)_T$ and is called a symmetric triangular fuzzy number.

Algebraic operations on fuzzy numbers are defined based on the extension principle. Here we recall this principle and also two well-known results [26].

Definition 5.3. (Extension Principle). Let X be the Cartesian product of universes $X_1 \times ... \times X_n$ and $\widetilde{A}_1,\ \widetilde{A}_2, ..., \widetilde{A}_n$ be n fuzzy sets in $X_1, ..., X_n$, respectively. Suppose that f is a mapping from X to a universe Y, $y = f(x_1, ..., x_n)$. Then, the extension principle allows us to define a fuzzy set $\widetilde{B}$ in Y by

$$\widetilde{B} = \{(y,\ \widetilde{B}(y)) | y = g(x_1, x_2, ..., x_n),\ (x_1, x_2, ..., x_n) \in X\}$$

where,

$$\widetilde{B}(y) = \begin{cases} \sup_{(x_1,x_2,...,x_n) \in f^{-1}(y)} min\{\widetilde{A}_1(x_1), ..., \widetilde{A}_n(x_n)\}, & f^{-1}(y) \neq \phi, \\ \\ 0, & otherwise. \end{cases} \quad \forall y \in Y.$$

in which, $f^{-1}(y)$ is the inverse image of y, i.e. $f^{-1}(y) := \{\omega \in X : f(\omega) = y\}$.

Proposition 5.4. *Let $\widetilde{M} = (m, \alpha, \beta)_{LR}$ be a LR-type fuzzy number and $\lambda \in \Re$. Then,*

$$\lambda\widetilde{M} = \begin{cases} (\lambda m, \lambda\alpha, \lambda\beta)_{LR}, & \lambda > 0, \\ (\lambda m, -\lambda\beta, -\lambda\alpha)_{RL}, & \lambda < 0. \end{cases}$$

Proposition 5.5. *Let $\widetilde{M} = (m, \alpha, \beta)_{LR}$ and $\widetilde{N} = (n, \gamma, \delta)_{LR}$ be two LR-type fuzzy numbers. Then,*
$\widetilde{M} + \widetilde{N} = (m, \alpha, \beta)_{LR} + (n, \gamma, \delta)_{LR} = (m + n, \alpha + \gamma, \beta + \delta)_{LR}$.

There are several defuzzification methods introduced in literature. A common method which is used in this paper is as follows.

Definition 5.6. (Center of Gravity Defuzzification Method) [24]. Let $\widetilde{W}$ be a fuzzy set of $\Re$, then the defuzzified version of $\widetilde{W}$ is defined as a real number as follows:

$$def_{CoG}(\widetilde{W}) = \frac{\int_x x\widetilde{W}(x)dx}{\int_x \widetilde{W}(x)dx}.$$

Acknowledgements. The authors are grateful to E. Falah, A. Ariya, and M. Mosallaei from Namazee hospital of Shiraz for their valuable comments on gathering the data and defining the membership degrees for fuzzy diabetic cases. Also, we would like to thank N. Shokrpour at Center for Development of Clinical Research of Namazee Hospital for editorial assistance. This work was financially supported by a grant (87-4110) from Shiraz University of Medical Sciences Research Council. The authors are grateful to the referees whose comments greatly improved this paper.

References

[1] A. Agresti, *Categorical data analysis*, Wiley, New york, 2002.
[2] A. R. Arabpour and M. Tata, *Estimating the parameters of a fuzzy linear regression model*, Iranian Journal of Fuzzy Systems, **5** (2008), 1-19.
[3] S. C. Bagley, H. White and B. A. Golomb, *Logistic regression in the medical literature: standards for use and reporting with particular attention to one medical domain*, Journal of Clinical Epidemiology, **54** (2001), 979-985.
[4] A. Celmins, *Least squares model fitting to fuzzy vector data*, Fuzzy Sets and Systems, **22** (1987), 260-269.
[5] R. Coppi, P. D'Urso, P. Giordani and A. Santoro, *Least squares estimation of a linear regression model with LR fuzzy response*, Computational Statistics and Data Analysis, **51** (2006), 267-286.
[6] P. Diamond, *Least squares fitting of several fuzzy variables*, Proc. of the Second IFSA Congress, Tokyo, (1987), 20-25.
[7] D. Dubois, E. Kerre, R. Mesiar and H. Prade, *Fuzzy interval analysis, In: D. Dubois, H. Prade, eds.*, Fundamentals of Fuzzy Sets, Kluwer, 2000.
[8] A. S. Fauci, E. Braunwald, D. L. Kasper, S. L. Hauser, D. L. Longo, J. L. Jameson and J. Loscalzo, *Harrison's principals of internal medicine*, Wiley, New York, **II** (2008), 2275-2279.
[9] GAMS (General Algebraic Modeling System), *A high-level modeling system for mathematical programming and optimization*, GAMS Development Corporation, Washington, DC, USA, 2007.

[10] H. Hassanpour, H. R. Maleki and M. A. Yaghoobi, *A note on evaluation of fuzzy linear regression models by comparing membership functions*, Iranian Journal of Fuzzy Systems, **6** (2009), 1-6.
[11] D. H. Hong, J. Song and H. Y. Do, *Fuzzy least-squares linear regression analysis using shape preserving operation*, Information Sciences, **138** (2001), 185-193.
[12] LINGO 8.0, *A linear programming, integer programming, nonlinear programming and global optimization solver*, Lindo System Inc, 1415 North Dayton Str., Chicago, 2003.
[13] MATLAB R., *A technical computing environment for high-performance numeric computation and Visualization*, The Math Works Inc., 2007.
[14] S. Mirzaei Yeganeh and S. M. Taheri, *Possibilistic logistic regression by linear programming approach*, Proc. of the 7th Seminar on Probability and Stochastic Processes, Isfahan University of Technology, Isfahan, Iran, (2009), 139-143.
[15] J. Mohammadi and S. M. Taheri, *Pedomodels fitting with fuzzy least squares regression*, Iranian Journal of Fuzzy Systems, **1** (2004), 45-61.
[16] G. Peters, *A linear forecasting model and its application in economic data*, Journal of Forecasting, **20** (2001), 315-328.
[17] S. Roychowdhury and W. Pedrycz, *Modeling temporal functions with granular regression and fuzzy rule*, Fuzzy Sets and Systems, **126** (2002), 377-387.
[18] B. Sadeghpour and D. Gien, *A goodness of fit index to reliability analysis in fuzzy model, In: A. Grmela, ed., Advances in Intelligent Systems, Fuzzy Systems, Evolutionary Computation*, WSEAS Press, Greece, (2002), 78-83.
[19] A. F. Shapiro, *Fuzzy regression models*, ARC, 2005.
[20] B. D. Tabaei, and W. H. Herman, *A multivariate logistic regression equation to screen for Diabetes*, Diabetes Care, **25** (2002), 1999-2003.
[21] S. M. Taheri, *Trends in fuzzy statistics*, Austrian Journal of Statistics, **32** (2003), 239-257.
[22] S. M. Taheri and M. Kelkinnama, *Fuzzy least absolutes regression*, Proc. of 4th International IEEE Conference on Intelligent Systems, Varna, Bulgaria, **11** (2008), 55-58.
[23] H. Tanaka, S. Uejima, K. Asai, *Linear regression analysis with fuzzy model*, IEEE Trans. Systems Man Cybernet., **12** (1982), 903-907.
[24] E. Van Broekhoven and B. D. Baets, *Fast and accurate of gravity defuzzification of fuzzy systems outputs defined on trapezoidsal fuzzy partitions*, Fuzzy Sets and Systems, **157** (2006), 904-918.
[25] L. A. Zadeh, *Fuzzy sets*, Information and Control, **8** (1965), 338-353.
[26] H. J. Zimmermann, *Fuzzy set theory and its applications*, 3rd ed., Kluwer, Dodrecht, 1996.

Saeedeh Pourahmad, Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, 71345-1874, Iran
E-mail address: pourahmad@sums.ac.ir

S. Mohammad Taghi Ayatollahi*, Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, 71345-1874, Iran
E-mail address: ayatolahim@sums.ac.ir

S. Mahmoud Taheri, Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, 84156-83111, Iran
E-mail address: taheri@cc.iut.ac.ir

*Corresponding author